

An Asymptotically Optimal Algorithm for the Max k -Armed Bandit Problem

Matthew J. Streeter¹ and Stephen F. Smith²

Computer Science Department

and Center for the Neural Basis of Cognition¹ and

The Robotics Institute²

Carnegie Mellon University

Pittsburgh, PA 15213

{matts, sfs}@cs.cmu.edu

Abstract

We present an asymptotically optimal algorithm for the *max* variant of the k -armed bandit problem. Given a set of k slot machines, each yielding payoff from a fixed (but unknown) distribution, we wish to allocate trials to the machines so as to maximize the expected maximum payoff received over a series of n trials. Subject to certain distributional assumptions, we show that $O\left(k \ln\left(\frac{k}{\delta}\right) \frac{\ln(n)^2}{\epsilon^2}\right)$ trials are sufficient to identify, with probability at least $1 - \delta$, a machine whose expected maximum payoff is within ϵ of optimal. This result leads to a strategy for solving the problem that is asymptotically optimal in the following sense: the gap between the expected maximum payoff obtained by using our strategy for n trials and that obtained by pulling the single best arm for all n trials approaches zero as $n \rightarrow \infty$.

1. Introduction

In the k -armed bandit problem one is faced with a set of k slot machines, each having an arm that, when pulled, yields a payoff from a fixed (but unknown) distribution. The goal is to allocate trials to the arms so as to maximize the expected cumulative payoff obtained over a series of n trials. Solving the problem entails striking a balance between exploration (determining which arm yields the highest mean payoff) and exploitation (repeatedly pulling this arm).

In the max k -armed bandit problem, the goal is to maximize the expected *maximum* (rather than cumulative) payoff. This version of the problem arises in practice when tackling combinatorial optimization problems for which a number of randomized search heuristics exist: given k heuristics, each yielding a stochastic outcome when applied to some particular problem instance, we wish to allocate trials to the heuristics so as to maximize the maximum payoff (e.g., the maximum number of clauses satisfied by any sampled variable assignment, the minimum makespan of any sampled schedule). Cicirello and Smith (2005) show that a max k -armed bandit approach yields good performance on the resource-constrained project scheduling problem with maximum time lags (RCPSP/max).

Copyright © 2006, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

1.1. Summary of Results

We consider a restricted version of the max k -armed bandit problem in which each arm yields payoff drawn from a *generalized extreme value (GEV) distribution* (defined in §2). This paper presents the first provably asymptotically optimal algorithm for this problem.

Roughly speaking, the reason for assuming a GEV distribution is the Extremal Types Theorem (stated in §2), which states that the distribution of the sample maximum of n independent identically distributed random variables approaches a GEV distribution as $n \rightarrow \infty$. A more formal justification is given in §3. For reasons that will become clear, the nature of our results depends on the shape parameter (ξ) of the GEV distribution. Assuming all arms have $\xi \leq 0$, our results can be summarized as follows.

- Let a be an arm that yields payoff drawn from a GEV distribution with unknown parameters; let M_n denote the maximum payoff obtained after pulling a n times; and let $m_n = \mathbb{E}[M_n]$. We provide an algorithm that, after pulling the arm $O\left(\ln\left(\frac{1}{\delta}\right) \frac{\ln(n)^2}{\epsilon^2}\right)$ times, produces an estimate \bar{m}_n of m_n with the property that $\mathbb{P}[|\bar{m}_n - m_n| < \epsilon] \geq 1 - \delta$.
- Let a_1, a_2, \dots, a_k be k arms, each yielding payoff from (distinct) GEV distributions with unknown parameters. Let m_n^i denote the expected maximum payoff obtained by pulling the i^{th} arm n times, and let $m_n^* = \max_{1 \leq i \leq k} m_n^i$. We provide an algorithm that, when run for n pulls, obtains expected maximum payoff $m_n^* - o(1)$.

Our results for the case $\xi > 0$ are similar, except that our estimates and expected maximum payoffs come within arbitrarily small *factors* (rather than absolute distances) of optimality. Specifically, our estimates have the property that $\mathbb{P}\left[\frac{1}{1+\epsilon} < \frac{\bar{m}_n - \alpha_1}{m_n - \alpha_1} < 1 + \epsilon\right] \geq 1 - \delta$ for constant α_1 independent of n , while the expected maximum payoff obtained by using our algorithm for n pulls is $m_n^*(1 - o(1))$.

1.2. Related Work

The classical k -armed bandit problem was first studied by Robbins (1952) and has since been the subject of numerous papers; see Berry and Fristedt (1986) and Kaelbling (1993) for overviews. In a paper similar in spirit to ours, Fong (1995) showed that an initial exploration phase consisting

of $O\left(\frac{k}{\epsilon^2} \ln\left(\frac{k}{\delta}\right)\right)$ pulls is sufficient to identify, with probability at least $1 - \delta$, an arm whose mean payoff is within ϵ of optimal. Theorem 2 of this paper proves a bound similar to Fong's on the number of pulls needed to identify an arm whose expected *maximum* payoff (over a series of n trials) is near-optimal.

The max variant of the k -armed bandit problem was first studied by Cicirello and Smith (2004; 2005), who successfully used a heuristic for the max k -armed bandit problem to select among priority rules for the RCPSP/max. The design of Cicirello and Smith's heuristic is motivated by an analysis of the special case in which each arm's payoff distribution is a GEV distribution with shape parameter $\xi = 0$, but they do not rigorously analyze the heuristic's behavior. Our paper is more theoretical and less empirical: on the one hand we do not perform experiments on any practical combinatorial problem, but on the other hand we provide stronger performance guarantees under weaker distributional assumptions.

1.3. Notation

For an arbitrary cumulative distribution function G , let the random variable M_n^G be defined by

$$M_n^G = \max\{Z_1, Z_2, \dots, Z_n\}$$

where Z_1, Z_2, \dots, Z_n are independent random variables, each having distribution G . Let

$$m_n^G = \mathbb{E}[M_n^G].$$

2. Extreme Value Theory

This section provides a self-contained overview of results in extreme value theory that are relevant to this work. Our presentation is based on the text by Coles (2001).

The central result of extreme value theory is an analogue of the central limit theorem that applies to extremely rare events. Recall that the central limit theorem states that (under certain regularity conditions) the distribution of the sum of n independent, identically distributed (i.i.d) random variables converges to a normal distribution as $n \rightarrow \infty$. The extremal types theorem states that (under certain regularity conditions) the distribution of the maximum of n i.i.d random variables converges to a generalized extreme value (GEV) distribution.

Definition (GEV distribution). A random variable Z has a generalized extreme value distribution if, for constants μ , $\sigma > 0$, and ξ , $\mathbb{P}[Z \leq z] = GEV_{(\mu, \sigma, \xi)}(z)$, where

$$GEV_{(\mu, \sigma, \xi)}(z) = \exp\left(-\left(1 + \frac{\xi(z - \mu)}{\sigma}\right)^{-\frac{1}{\xi}}\right)$$

for $z \in \{z : 1 + \xi(z - \mu)\sigma^{-1} > 0\}$, and $GEV_{(\mu, \sigma, \xi)}(z) = 1$ otherwise. The case $\xi = 0$ is interpreted as the limit

$$\lim_{\xi' \rightarrow 0} GEV_{(\mu, \sigma, \xi')}(z) = \exp\left(-\exp\left(\frac{\mu - z}{\sigma}\right)\right).$$

The following three propositions establish properties of the GEV distribution.

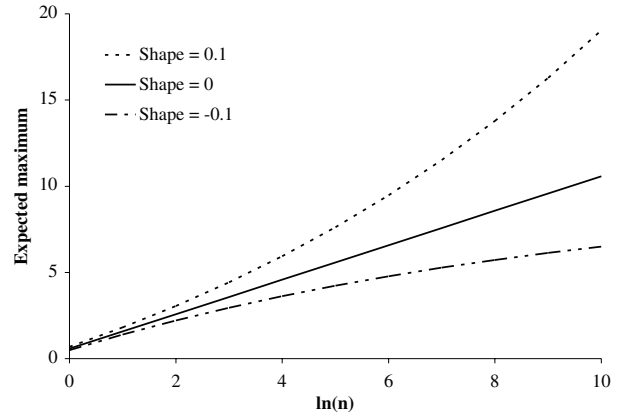


Figure 1: The effect of the shape parameter (ξ) on the expected maximum of n independent draws from a GEV distribution.

Proposition 1. Let Z be a random variable with $\mathbb{P}[Z \leq z] = GEV_{(\mu, \sigma, \xi)}(z)$. Then

$$\mathbb{E}[Z] = \begin{cases} \mu + \frac{\sigma}{\xi} (\Gamma(1 - \xi) - 1) & \text{if } \xi < 1, \xi \neq 0 \\ \mu + \sigma\gamma & \text{if } \xi = 0 \\ \infty & \text{if } \xi \geq 1 \end{cases}$$

where

$$\Gamma(z) = \int_0^\infty t^{z-1} \exp(-t) dt$$

is the complete gamma function and

$$\gamma = \lim_{n \rightarrow \infty} \left(\sum_{k=1}^n \frac{1}{k} - \ln(n) \right)$$

is Euler's constant.

Proposition 2. Let $G = GEV_{(\mu, \sigma, \xi)}$. Then M_n^G has distribution $GEV_{(\mu', \sigma', \xi')}$, where

$$\begin{aligned} \mu' &= \begin{cases} \mu + \frac{\sigma}{\xi} (n^\xi - 1) & \text{if } \xi \neq 0 \\ \mu + \sigma \ln(n) & \text{otherwise,} \end{cases} \\ \sigma' &= \sigma n^\xi, \text{ and} \\ \xi' &= \xi. \end{aligned}$$

Substituting the parameters of M_n^G given by Proposition 2 into Proposition 1 gives an expression for m_n^G .

Proposition 3. Let $G = GEV_{(\mu, \sigma, \xi)}$ where $\xi < 1$. Then

$$m_n^G = \begin{cases} \mu + \frac{\sigma}{\xi} (n^\xi \Gamma(1 - \xi) - 1) & \text{if } \xi \neq 0 \\ \mu + \sigma\gamma + \sigma \ln(n) & \text{otherwise.} \end{cases}$$

It follows that

- for $\xi > 0$, m_n^G is $\Theta(n^\xi)$;
- for $\xi = 0$, m_n^G is $\Theta(\ln(n))$; and
- for $\xi < 0$, $m_n^G = \mu - \frac{\sigma}{\xi} - \Theta(n^\xi)$.

It is useful to have a visual picture of what Proposition 3 means. Figure 1 plots m_n^G as a function of n for three GEV distributions with $\mu = 0$, $\sigma = 1$, and $\xi \in \{0.1, 0, -0.1\}$.

The central result of extreme value theory is the following theorem.

The Extremal Types Theorem. Let G be an arbitrary cumulative distribution function, and suppose there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\lim_{n \rightarrow \infty} \mathbb{P} \left[\frac{M_n^G - b_n}{a_n} \leq z \right] = G^*(z) \quad (1)$$

for any continuity point z of G^* , where G^* is a not a point mass. Then there exist constants $\mu, \sigma > 0$, and ξ such that $G^*(z) = GEV_{(\mu, \sigma, \xi)}(z) \forall z$. Furthermore,

$$\lim_{n \rightarrow \infty} \mathbb{P} [M_n \leq z] = GEV_{(\mu a_n + b_n, \sigma a_n, \xi)}(z).$$

Condition (1) holds for a variety of distributions including the normal, lognormal, uniform, and Cauchy distributions.

3. The Max k -Armed Bandit Problem

Definition (max k -armed bandit instance). An instance $I = (n, \mathcal{G})$ of the max k -armed bandit problem is an ordered pair whose first element is a positive integer n , and whose second element is a k -tuple $\mathcal{G} = (G_1, G_2, \dots, G_k)$ of cumulative distribution functions, each thought of as an arm on a slot machine. The i^{th} arm, when pulled, returns a sample drawn independently at random from G_i .

Definition (max k -armed bandit strategy). A max k -armed bandit strategy S is an algorithm that, given an instance $I = (n, \mathcal{G})$ of the max k -armed bandit problem, performs a sequence of n arm-pulls. For any strategy S and integer $\ell \leq n$, we denote by $S_\ell(I)$ the expected maximum payoff obtained by running S on I for ℓ trials:

$$S_\ell(I) = \mathbb{E} \left[\max_{0 \leq j \leq \ell} p_j \right]$$

where p_j is the payoff obtained from the j^{th} pull, and we define $p_0 = 0$.

Our goal is to come up with a strategy S such that $S_n(I)$ is near-maximal.

Note that the problem is ill-posed (i.e., there is no clear criterion for preferring one strategy over another) unless we make some assumptions about the distributions G_i . We will assume that each arm $G_i = GEV_{(\mu_i, \sigma_i, \xi_i)}$ is a GEV distribution whose parameters satisfy

1. $|\mu_i| \leq \mu_u$
2. $0 < \sigma_\ell \leq \sigma_i \leq \sigma_u$
3. $\xi_\ell \leq \xi_i \leq \xi_u < \frac{1}{2}$

for known constants $\mu_u, \sigma_\ell, \sigma_u, \xi_\ell$, and ξ_u .

There are two arguments for assuming that each arm is a GEV distribution. First, in practice the distribution of payoffs returned by a strong heuristic may be approximately GEV, even if the conditions of the Extremal Type Theorem are not formally satisfied (Cicirello & Smith 2004).

A second argument runs as follows. Suppose $I = (n, \mathcal{G})$ is an instance of the max k -armed bandit problem in which each distribution $G_i \in \mathcal{G}$ satisfies condition (1) of the Extremal Types Theorem. Consider the instance $\bar{I} = (\frac{n}{m}, \bar{\mathcal{G}})$, where $\bar{\mathcal{G}} = (\bar{G}_1, \bar{G}_2, \dots, \bar{G}_k)$, and arm \bar{G}_i returns the maximum payoff obtained by pulling the corresponding arm G_i

m times. Effectively, \bar{I} is a restricted version of I in which the arms must be pulled in batches of size m , rather than in any arbitrary order. For m sufficiently large, the Extremal Types Theorem guarantees that for each i , $\bar{G}_i \approx GEV_{(\mu_i, \sigma_i, \xi_i)}$ for some constants μ_i, σ_i , and ξ_i . Thus, the instance $I' = (\frac{n}{m}, \mathcal{G}')$ with $\mathcal{G}' = (G'_1, G'_2, \dots, G'_k)$ and $G'_i = GEV_{(\mu_i, \sigma_i, \xi_i)}$ is approximately equivalent to \bar{I} and satisfies our distributional assumptions.

The purpose of the restrictions on the parameters μ_i, σ_i , and ξ_i is to ensure that each GEV distribution has finite, bounded mean and variance.

4. An Asymptotically Optimal Algorithm

In this section we will analyze the max k -armed bandit strategy \mathcal{S}^1 shown below.

Our analysis will take a different form depending on whether each GEV distribution has shape parameter $\xi < 0$, $\xi = 0$, or $\xi > 0$. Although we will analyze all three cases, it is worth noting that the case $\xi < 0$ is the only one that can arise in practice. This is true because in any real combinatorial optimization problem the maximum payoff is bounded from above, which (by Proposition 3) can only happen when $\xi < 0$.

Strategy $\mathcal{S}^1(\epsilon, \delta)$:

1. (*Exploration*) For each arm $G_i \in \mathcal{G}$:

Using $t = O\left(\ln\left(\frac{1}{\delta}\right) \frac{\ln(n)^2}{\epsilon^2}\right)$ samples of G_i , obtain an estimate $\bar{m}_n^{G_i}$ of $m_n^{G_i}$. Assuming that arm G_i has shape parameter $\xi_i \leq 0$, our estimate will have the property that

$$\mathbb{P} [|\bar{m}_n^{G_i} - m_n^{G_i}| < \epsilon] \geq 1 - \delta.$$

2. (*Exploitation*) Set $\hat{i} = \arg \max_{1 \leq i \leq k} \bar{m}_n^{G_i}$, and pull arm $G_{\hat{i}}$ for the remaining $n - tk$ trials.

If an arm G_i has shape parameter $\xi_i > 0$, the estimate obtained in step 1 (a) will instead have the property that $\mathbb{P} \left[\frac{1}{1+\epsilon} < \frac{\bar{m}_n - \alpha_1}{m_n - \alpha_1} < 1 + \epsilon \right] \geq 1 - \delta$ for constant α_1 independent of n .

The following theorem shows that with appropriate settings of ϵ and δ , strategy \mathcal{S}^1 is asymptotically optimal when each arm has shape parameter $\xi_i \leq 0$. In Appendix A, we establish a similar guarantee (using the same parameter settings) when one or more arms have $\xi_i > 0$.

Theorem 1. Let $I = (n, \mathcal{G})$ be an instance of the max k -armed bandit problem, where $\mathcal{G} = (G_1, G_2, \dots, G_k)$ and $G_i = GEV_{(\mu_i, \sigma_i, \xi_i)}$. Let

- $m_n^* = \max_{1 \leq i \leq k} m_n^{G_i}$,
- $\xi_{max} = \max_{1 \leq i \leq k} \xi_i$, and
- $S = \mathcal{S}^1 \left(\sqrt[3]{\frac{k}{n}}, \frac{1}{kn^2} \right)$.

Then if $\xi_{max} \leq 0$,

$$S_n(I) = m_n^* - O(\Delta)$$

where $\Delta = \ln(nk) \ln(n)^2 \sqrt[3]{\frac{k}{n}}$.

Proof. Let $\hat{m}_n = m_n^{G_{\hat{i}}}$ (where \hat{i} is the arm selected for exploitation in step 2). Then \hat{m}_{n-tk} is the expected maximum payoff obtained during the exploitation step, so

$$S_n(I) \geq \hat{m}_{n-tk}.$$

Claim 1. $\hat{m}_n - \hat{m}_{n-tk}$ is $O(\frac{tk}{n})$.

Proof of Claim 1. Let $\mu = \mu_{\hat{i}}$, $\sigma = \sigma_{\hat{i}}$, and $\xi = \xi_{\hat{i}}$ be the parameters of the arm selected for exploitation. Suppose $\xi = 0$. Then by Proposition 3, $\hat{m}_n - \hat{m}_{n-tk} = \sigma(\ln(n) - \ln(n-tk))$. Thus for n sufficiently large,

$$\begin{aligned} \hat{m}_n - \hat{m}_{n-tk} &= \sigma(\ln(n) - \ln(n-tk)) \\ &= -\sigma \ln\left(\frac{n-tk}{n}\right) \\ &= -\sigma \ln\left(1 - \frac{tk}{n}\right) \\ &< 2\sigma \frac{tk}{n} \\ &= O\left(\frac{tk}{n}\right) \end{aligned}$$

where on the fourth line we have used the fact that for n sufficiently large, $\frac{tk}{n} < \frac{1}{2}$, and for $0 < x < \frac{1}{2}$, $-2x < \ln(1-x) \leq -x$.

Now suppose $\xi < 0$. By Proposition 3, $\hat{m}_n - \hat{m}_{n-tk} = \frac{\sigma}{\xi} \Gamma(1-\xi)(n^\xi - (n-t)^\xi) = O((n-t)^\xi - n^\xi)$ where we have used the fact that $\frac{\sigma}{\xi} \Gamma(1-\xi) < 0$. Expanding $(n-t)^\xi$ in powers of t about $t=0$ gives

$$(n-t)^\xi = n^\xi - \xi n^{\xi-1}t + O(t^2 n^{\xi-2}).$$

Because $\xi \leq 0$ and $|\xi|$ is bounded, it follows that $(n-t)^\xi - n^\xi$ is $O(\frac{t}{n})$. \square

With probability at least $1 - k\delta$, all estimates obtained during the exploration phase are within ϵ of the correct values, so that $m_n^* - \hat{m}_n < 2\epsilon$. Assuming $m_n^* - \hat{m}_n < 2\epsilon$, it follows that

$$\begin{aligned} m_n^* - \hat{m}_{n-tk} &= (m_n^* - \hat{m}_n) + (\hat{m}_n - \hat{m}_{n-tk}) \\ &< 2\epsilon + O\left(\frac{tk}{n}\right) \\ &= 2\epsilon + \frac{k}{n} O\left(\ln\left(\frac{1}{\delta}\right) \frac{(\ln n)^2}{\epsilon^2}\right) \\ &= O(\Delta) \end{aligned}$$

where on the second line we have used Claim 1. Thus with probability at least $1 - k\delta$, our expected maximum payoff is at least $m_n^* - O(\Delta)$. Therefore,

$$\begin{aligned} S_n(I) &\geq (1 - k\delta)(m_n^* - O(\Delta)) \\ &\geq m_n^* - O(\Delta) - k\delta m_n^* \\ &= m_n^* - o(1) \end{aligned}$$

where on the last line we have used the fact that $\Delta = o(1)$ and the fact that for $\xi \leq 0$, m_n^* is $O(\log n)$, so that $k\delta m_n^* = \frac{m_n^*}{n^2} = o(1)$. \square

Theorem 1 completes our analysis of the performance of S^1 . It remains only to describe how the estimates in step 1 (a) are obtained.

4.1. Estimating m_n

We adopt the following notation:

- Let $G = GEV_{\mu, \sigma, \xi}$ denote a GEV distribution with (unknown) parameters μ , σ , and ξ satisfying the conditions stated in §3, and
- let $m_i = m_i^G$.

To estimate m_n , we first obtain an accurate estimate of ξ . Then

1. if $\xi \approx 0$ (so that the growth of m_n as a function of $\ln n$ is linear), we estimate m_n by first estimating m_1 and m_2 , then performing linear interpolation;
2. otherwise we estimate m_n by first estimating m_1, m_2 , and m_4 , then performing a nonlinear interpolation.

4.1.1. Estimating m_i for $i \in \{1, 2, 4\}$

The following two lemmas use well-known ideas to efficiently estimate m_i for small values of i .

Lemma 1. For any fixed positive integer i , $O(\frac{1}{\epsilon^2})$ draws from G suffice to obtain an estimate \bar{m}_i of m_i such that

$$\mathbb{P}[|\bar{m}_i - m_i| < \epsilon] \geq \frac{3}{4}.$$

Proof. First consider the special case $i=1$. Let X denote the sum of t draws from G , for some to-be-specified positive integer t . Then $\mathbb{E}[X] = m_1 t$ and $\text{Var}[X] = \tilde{\sigma}^2 t$, where $\tilde{\sigma}$ is the (unknown) standard deviation of G ($\tilde{\sigma}$ is proportional to, but not the same as, the scale parameter σ). We take $\bar{m}_1 = \frac{X}{t}$ as our estimate of m_1 . Then

$$\begin{aligned} \mathbb{P}[|\bar{m}_1 - m_1| \geq \epsilon] &= \mathbb{P}[|t\bar{m}_1 - tm_1| \geq t\epsilon] \\ &= \mathbb{P}[|X - \mathbb{E}[X]| \geq \frac{\sqrt{t}\epsilon}{\tilde{\sigma}} \sqrt{\text{Var}[X]}] \\ &\leq \frac{\tilde{\sigma}^2}{t\epsilon^2} \end{aligned}$$

where the last inequality is Chebyshev's. Thus to guarantee $\mathbb{P}[|\bar{m}_1 - m_1| \geq \epsilon] \leq \frac{1}{4}$ we must set $t = \frac{4\tilde{\sigma}^2}{\epsilon^2} = O(\frac{1}{\epsilon^2})$ (note that due to the assumptions in §3, $\tilde{\sigma}$ is $O(1)$).

For $i > 1$, we let X be the sum of t block maxima (each the maximum of i independent draws from G). Because the standard deviation of M_i and i itself are both $O(1)$, the lemma follows. \square

To boost the probability that $|\bar{m}_i - m_i| < \epsilon$ from $\frac{3}{4}$ to $1 - \delta$, we use the ‘‘median of means’’ method.

Lemma 2. Let i be a positive integer and let $\epsilon > 0$ and $\delta \in (0, 1)$ be real numbers. Then $O(\ln(\frac{1}{\delta}) \frac{i}{\epsilon^2})$ draws from G suffice to obtain an estimate \bar{m}_i of m_i such that

$$\mathbb{P}[|\bar{m}_i - m_i| < \epsilon] \geq 1 - \delta.$$

Proof. We invoke Lemma 1 r times (for r to be determined), yielding a set $E = \{\bar{m}_i^{(1)}, \bar{m}_i^{(2)}, \dots, \bar{m}_i^{(r)}\}$ of estimates of m_i . Let \bar{m}_i be the median element of E . Let $A = \{\bar{m}_i^{(j)} \in E : |\bar{m}_i^{(j)} - m_i| < \epsilon\}$ be the set of ‘‘accurate’’ estimates of m_i ; and let $A = |A|$. Then $|\bar{m}_i - m_i| \geq \epsilon$ implies $A \leq \frac{r}{2}$,

